Pràctica2 - M2.951\_20221

Gerard Ramos Gambús: [gambusrg@uoc.edu](mailto:gambusrg@uoc.edu)

Oriol Caravaca Müller: ocaravacam@uoc.edu

03/01/2023

Contingut

[Descripció del *dataset* 3](#_Toc124333004)

[Integració i selecció 4](#_Toc124333005)

[Neteja de les dades 4](#_Toc124333006)

[Anàlisi de les dades 4](#_Toc124333007)

[Anàlisi visual 4](#_Toc124333008)

[Normalitat i Homoscedasticitat 7](#_Toc124333009)

[Proves estadístiques 9](#_Toc124333010)

[Test d’hipòtesis 10](#_Toc124333011)

[Regressions 10](#_Toc124333012)

[Representació dels resultats 13](#_Toc124333013)

[Resolució del problema 13](#_Toc124333014)

[Codi 13](#_Toc124333015)

[Video 13](#_Toc124333016)

# Descripció del *dataset*

El *dataset* triat és “[Heart Attack Analysis & Prediction dataset](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)”, publicat per Rashik Rahman a [Kaggle](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset) amb llicencia CC0 1.0 Universal (CC0 1.0) *Public Domain Dedication.* Aquest *dataset* conte 13 variables corresponents a dades demogràfiques i resultats de proves mèdiques cardiovasculars de 303 pacients. Així mateix, ens ofereix una variable objectiu que ens indica quin es el risc de patir un infart de cada un d’aquest pacients.

Aquest *dataset* és d’interès perquè ens permet detectar factors de risc i crear models predictius que poden ajudar a la diagnosis i detecció de problemes cardiovasculars.

A continuació es detalla el tipus i descripció de les variables del *dataset*:

|  |  |  |
| --- | --- | --- |
| Variable | Tipus | Descripció |
| age | Categòrica | Edat de la persona |
| sex | Categòrica | Gènere de la persona |
| cp | Categòrica | Tipus de dolor toràcic  1= angina típica 2 = angina atípica 3 = dolor no anginós 4 = asimptomàtic |
| trtbps | Numèrica | Pressió arterial en repòs (en mm Hg) |
| chol | Numèrica | Colestorol en mg/dl obtingut mitjançant el sensor IMC |
| fbs | Categòrica | Sucre en sang en dejú > 120 mg/dl (1 = cert; 0 = fals) |
| restecg | Categòrica | Resultats electrocardiogràfics en repòs  0 = normal 1 = tenir una anomalia de l'ona ST-T 2 = mostra una hipertròfia ventricular esquerre probable o definitiva |
| thalachh | Numèrica | Freqüència cardíaca màxima aconseguida |
| exng | Categòrica | Angina induïda per l'exercici (1 = sí; 0 = no) |
| oldpeak | Numèrica | Depressió del ST induïda per l'exercici en relació amb el repòs |
| slp | Categòrica | El pendent del segment ST de l'exercici màxim 0 = sense pendent 1 = pla 2 = pendent avall |
| caa | Numèrica | Nombre de arteries principals (0-3) |
| thall | Categòrica | Talassèmia 0 = nul 1 = defecte fixat 2 = normal 3 = defecte reversible |
| output | Categòrica | 0 = menys probabilitat d'atac cardíac  1 = més probabilitat d'atac cardíac |

# Integració i selecció

Per a la realització d’aquesta practica es contempla la utilització del *dataset* complet amb l’objectiu de fer-ne un anàlisis i implementar un model de regressió logística.

De l’estudi del model de regressió logística se’n descarta la variable fbs per considerar-se insignificant amb un nivell de significació per sota del 95%.

# Neteja de les dades

El que s’ha realitzat per a la neteja de les dades ha estat fer comprovacions sobre tot el conjunt de dades, tenint en compte si hi havia espais en blanc en les observacions o valors NA. El que s’ha vist és que les dades son consistents i no tenen valors perduts.

D’altra banda sí que podem afirmar que el conjunt de dades conté valors atípics.

De les variables que més valors atípics contenien, el que s’ha fet és identificar-los i tractar-los utilitzant la funció kNN de R, la qual utilitza l’algorisme K-nearest neighbours per reomplir amb valors que tinguin sentit les observacions que son outliers.

# Anàlisi de les dades

## Anàlisi visual

En primer lloc el que s’ha fet és fer un anàlisi visual per comprendre com son i com estan relacionades les dades entre si. Per això s’han creat un conjunt de diagrames que faciliten aquesta comprensió.

En primer lloc s’han creat diagrames de sectors, que representen el percentatge d’homes i dones de tot el conjunt de dades, el qual son un 68% homes i 32% dones,

Gráfico, Gráfico circular

Descripción generada automáticamente

un en funció del percentatge de pacients segons el risc d’infart, el qual determina que un 54% del pacients tenen probabilitats altes de patir un infart i el 46% baixes.

Gráfico, Gráfico circular

Descripción generada automáticamente

En segon lloc, s’han creat diagrames de barres per veure com es distribueixen les dades.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Atenent al gràfic, els homes tenen moltes més angines típiques que les dones, però de moment no podem fer cap suposició sobre que representi si les angines típiques tenen més per a l’hora de patir un infart.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Segons els resultats electrocardiogràfics, hi ha una alta probabilitat de tenir un atac al cor si hi ha anomalies de l'ona ST-T.

Seguidament s’ha fet una discretització per intervals (no iguals) de l’edat per determinar quin rang d’edats son els més propensos a patir infarts.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Les edats s’han separat en Jove/Adult (27-40 anys), Adult(41-55 anys), Gent gran(56-65 anys) i Avis(66-78 anys).

El que es pot extreure del gràfic és que les persones més propenses a tenir un atac de cor son les els adults. D’altra banda, dels que tenen baixa probabilitat de patir un infart, la gent gran és la que més possibilitats té de patir-lo.

## Normalitat i Homoscedasticitat

Per a cada variable continua s'aplica un test Shapiro-Wilk i se’n mostra la distribució amb un gràfic de densitat per comprovar-ne la normalitat. Només la variable 'chol' passa el test, afortunadament sabem que gracies el teorema del límit central si el nombre d'observacions es major que 30 les variables es poden tractar coma variables amb distribució normal.

Els resultats obtinguts han estat els següents:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

A simple vista no totes les variables segueixen una distribució normal, però s’assumeix que si degut als gràfics i al TLC, el qual determina que si la mida de la mostra és superior a 30 elements i les mitjanes de les mostres s’aproximen a la mitjana de la població, s’aproxima a una distribució normal.

///HOMOSCEDASTICITAT

Per cada variable continua es comprova la Homoscedasticitat entre la seva distribució en relació a la probabilitat de patir un infart i al sexe.

## Proves estadístiques

Per realitzar les proves estadístiques s’han aplicat:

* Test d’hipòtesis
* Regressió lineal
* Correlacions

Per aplicar el test d’hipòtesis s’han desenvolupat tres possibles hipòtesis.

### Test d’hipòtesis

**1-El sexe influeix en el risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

* H0 : No existeix correlació entre les variables.
* H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s’ha obtingut **p-value = 1.007e-06**. Com que p < 0.05, rebutgem la hipòtesis nul·la.

El sexe és un component que afecta al risc d'infart.

**2-L'edat afecta al risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

* H0 : No existeix correlació entre les variables.
* H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s’ha obtingut **p-value = 0.1309**. Com que p > 0.05, acceptem la hipòtesis nul·la.

L’edat no és un component que afecta al risc d'infart.

**3-El colesterol en sang afecta al risc de infart?.**

**Hipòtesi nul·la i l'alternativa**

* H0 : No existeix correlacio entre les variable
* H1 : Existeixen correlació entre les variables.

Aplicant el Chi-square test, s’ha obtingut **p-value = 0.07738**. Com que p > 0.05, acceptem la hipòtesis nul·la.

El colesterol en sang no és un component que afecta al risc d'infart.

### Regressions

Per fer l’anàlisi de regressió, s’ha creat un primer model.

Ss’ha dividit el conjunt d’entrenament en train i test, amb un 80/20 del total de les dades respectivament. Posteriorment s’ha creat un model **glm**(Generalized Linear Model) el qual hem determinat que la variable output seria la dependent, i age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall les variables independents del model.

Al fer un summary del model hem obtingut els següents resultats.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Veient els valor de p, i els significance codes, podem determinar que les variables explicatives o independents estadísticament més significants son cp2, caa1 i caa2 com les més significatives, ja que els seus valors estan en el rang [0, 0.001].

Després s’ha realitzat una anàlisi de col·linealitat del model amb una matriu de correlacions i amb la funció VIF de R.

Calendario

Descripción generada automáticamente con confianza media

Un texto con letras negras

Descripción generada automáticamente

Com es pot comprovar, les dues formes demostren que no hi ha una col·linealitat significativa entre les variables.

Seguidament s’ha creat el model final excloent les variables que no eren significatives.

Texto

Descripción generada automáticamente

//Correlacions

# Representació dels resultats

# Resolució del problema

# Codi

# Video

|  |  |
| --- | --- |
| **Contribucions** | **Signatura** |
| Investigació prèvia |  |
| Redacció de les respostes |  |
| Desenvolupament del codi |  |
| Participació al vídeo |  |